Technical Report 798

Review of Command Group Training Measurement Methods

Delane K. Garlinger and Jon J. Fallesen

ARI Field Unit at Fort Leavenworth, Kansas
Systems Research Laboratory

U. S. Army
Research Institute for the Behavioral and Social Sciences

July 1988

U. S. ARMY RESEARCH INSTITUTE
FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the

Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Technical Director

WM. DARRYL HENDERSON
COL, IN
Commanding

Technical review by

Dee Andrews
Patrick J. Whitmarsh

NOTICES

DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-POT, 5001 Eisenhower Ave., Alexandria, Virginia 22333-5600.

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) ARI Technical Report 798	5. MONITORING ORGANIZATION REPORT NUMBER(S) --
6a. NAME OF PERFORMING ORGANIZATION U.S. Army Research Institute 6b. OFFICE SYMBOL (If applicable) PERI-SL	7a. NAME OF MONITORING ORGANIZATION --
6c. ADDRESS (City, State, and ZIP Code) ARI Field Unit--Leavenworth P.O. Box 3407 Fort Leavenworth, KS 66027-0347	7b. ADDRESS (City, State, and ZIP Code) --
8a. NAME OF FUNDING/SPONSORING ORGANIZATION U.S. Army Research Institute 8b. OFFICE SYMBOL (If applicable) PERI-SZ	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER --
8c. ADDRESS (City, State, and ZIP Code) 5001 Eisenhower Avenue Alexandria, VA 22333-5600	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. 6.27.22.A PROJECT NO. 2Q162 722A791 TASK NO. 1.3.3. WORK UNIT ACCESSION NO. H.1

11. TITLE (Include Security Classification)
Review of Command Group Training Measurement Methods

12. PERSONAL AUTHOR(S)
Delane K. Garlinger and Jon J. Fallesen

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Final	FROM 03/86 TO 01/87	1988, July	48

16. SUPPLEMENTARY NOTATION
The reader is referred to ARI Research Report 1459, entitled "Feedback Principles for Command Group Training" by Delane K. Garlinger.

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Command group training; Training; Performance measurement; Feedback; Measurement techniques

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

This report documents a literature review of performance measurement for command group training, as well as providing a discussion of potential sources of performance data.

Specific measurement techniques (i.e., self-assessment, peer assessment, ARTEP, probes, battle outcome data, etc.), which have been applied and reported in the literature, are analyzed against 10 performance measurement criteria. Of those measurement techniques analyzed, none favorably met all 10 of the established criteria.

The analysis resulted in the determination that no one technique is acceptable in its present form for diagnosis and feedback in command group training, and that some combination of techniques, with refinements, will be needed. Several conclusions based upon the results are as follows:

(Continued)

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION
22a. NAME OF RESPONSIBLE INDIVIDUAL Delane K. Garlinger	22b. TELEPHONE (Include Area Code) (913) 684-4933 22c. OFFICE SYMBOL PERI-SL

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

ARI Technical Report 798

19. Abstract (Continued)

a. External observers are to be preferred over peer- or self-assessment.

b. Probes can enhance training exercises as well as present situations for measurement of subsequent performance.

c. Information flow and other testing techniques rate better than observation or summarization techniques in terms of objectivity, accuracy, validity, and reliability.

Areas identified for further research and development include better assessment of measurement techniques, especially in terms of validity, reliability, and accuracy; refinement of measures for staff perceptions, information usage, and secondary task performance; various uses of automated simulation and data tracking techniques; and better understanding of staff performance. (SDW)

Technical Report 798

Review of Command Group Training Measurement Methods

Delane K. Garlinger and Jon J. Fallesen

ARI Field Unit at Fort Leavenworth, Kansas
Stanley M. Halpin, Chief

Systems Research Laboratory
Robin L. Keesee, Director

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES

5001 Eisenhower Avenue, Alexandria, Virginia 22333-5600

Office, Deputy Chief of Staff for Personnel
Department of the Army

July 1988

Army Project Number
2Q162722A791

Manpower, Personnel, and Training

FOREWORD

The Fort Leavenworth Field Unit of the Army Research Institute for the Behavioral and Social Sciences supports the Combined Arms Center with research and development on combined arms operations and command group training. Measurement of staff performance is an issue common to research on both operations and training. In command group training, performance assessment is key for providing diagnostic feedback to the training audience.

This report provides a rigorous review of techniques that have been used to measure command group performance. Additional measurement techniques are discussed in terms of how they might be applied to command group training. The review documents the success, or lack of it, in developing command group performance measurement, and in doing so identifies several areas of needed research.

EDGAR M. JOHNSON
Technical Director

REVIEW OF COMMAND GROUP TRAINING MEASUREMENT METHODS

EXECUTIVE SUMMARY

Requirement:

The purpose of this report is to present a review of performance measurement methods by analyzing specific techniques that have been investigated and reported in the area of command group training (CGT), as well as presenting a discussion of potential classes and dimensions of performance that might provide diagnostic information for feedback purposes.

This report represents a preliminary step in a long-term effort to develop a set of procedural guidelines for tailoring diagnostic performance measures to a staff training exercise.

Procedure:

Based upon measurement theory and the constraints imposed by CGT environments, 10 criteria for selection and/or development of measurement techniques were established. The literature relevant to CGT measurement techniques was reviewed, and reported measures were assessed in terms of the 10 criteria. Other measurement methods that have not been applied in CGT, but that were considered to have potential for such an application, were described and assessed.

Findings:

Of those measurement techniques that have been tried previously in CGT and reviewed in this report, no overwhelming "success" was discovered. None of the techniques met all 10 of the desired measurement characteristics favorably. It is concluded that no one technique is acceptable in its present form for diagnostic feedback, and that some combination of techniques, with refinements, will be required.

Research and development efforts to produce the necessary refinements are recommended.

Utilization of Findings:

This report provides a comprehensive review of the current state of staff performance measurement that will be useful to the training community in specifying requirements, designing training systems, and evaluating performance, and to the larger community in evaluating command and control.

This analytical review of CGT performance measurement provides an initial step for developing the guidelines to help tailor measures for a particular training purpose and training event.

REVIEW OF COMMAND GROUP TRAINING MEASUREMENT METHODS

CONTENTS

LIST OF TABLES

REVIEW OF COMMAND GROUP TRAINING MEASUREMENT METHODS

INTRODUCTION

Army commanders and their staff groups must be capable of performing their command and control (C^2) functions at a high level of proficiency to ensure that the tenets of airland battle doctrine work. Staff groups train in a variety of modes such as command post exercises (CPX) and command field exercises (CFX). Training of corps and division staffs occurs only about twice a year because of the high costs incurred for high echelon training, and because of the time demands for these echelons to run garrison operations and to conduct training for subordinate units. Since the command group's proficiency in C^2 operations is so vital to battlefield performance and the opportunity for training is relatively infrequent, it is imperative to maximize the benefits derived from every training exercise. Research has consistently shown that objective performance feedback has a positive impact on subsequent performance (e.g., Downs, Johnson, & Barge, 1984; Ilgen, Fisher, & Taylor, 1984; and Thomas, Kaplan, & Barber, 1984).

Improved diagnostic feedback is needed for command group training (CGT). Feedback for CGT currently relies on an after action review process consisting of general observations about the events of the battle. The after action review session convenes at the conclusion of the training or at logical breakpoints, such as the end of the day. Diagnosis of performance is attempted by drawing out the participants' comments, but the staff typically gains little objective information concerning how well or how poorly their individual and collective tasks were performed.

As a prerequisite for providing objective feedback, "good" performance measurement is needed. By the nature of what command groups do, any thorough performance evaluation scheme quickly becomes complicated. Performance measurement for feedback purposes in CGT has been a recurrent concern. The Army Science Board in their report on Army training emphasized that the key to training improvement is performance measurement (1985).

This report provides a significant and encompassing review of the state of measurement for staff performance for the training community to use in requirements specification, training system design and evaluation, and for anyone involved in staff performance measurement.

This report also provides future directions for research and development work needed to produce a set of procedural guidelines for tailoring diagnostic measures to a staff training exercise. The resulting procedures would guide the trainer to derive a prioritization for what needs to be measured and how, given the resources available and the training objectives. The goal is to have guidelines available for use by the trainers to select and develop measurement

techniques to be incorporated into an exercise, to assist in preparation for executing the measurement plan, to prepare results for feedback and to guide in providing feedback. For these long-term goals, this report documents the advantages and disadvantages of measures and measurement techniques. This information is essential for selecting appropriate measurement techniques to match to CGT tasks.

In order to provide a structured and coherent framework within which to present the literature related to CGT training, sources and dimensions of CGT performance data have been categorized into general classes of performance measurment, and ten criteria have been used to describe and assess the performance measurement methods which were reviewed. The measurement criteria, sources and dimensions of performance, and classes of performance measurement are discussed more fully below.

Measurement Criteria

Since diagnostic assessment in CGT is dependent upon measurement of performance, it is crucial that the measures used be as dependable as possible. Although there will always be some error associated with performance measurement, the objective is to do everything feasible to limit error to a minimum. To this end, the following ten criteria for selection and/or development of measurement techniques have been set forth based upon current measurement theory (Anastasi, 1982; Thorndike, 1982) and the constraints imposed by CGT environments. Some of these criteria are interrelated, but each offers a concept which should be considered in the selection/development of performance measures.

Available for Timely Feedback

If measurement information is to be useful for feedback in training, it must be provided during or immediately following the completion of a training session, and must be presented in a form that is understood by the user in the context of training. Performance information loses a major part of its instructional value if it is not available quickly and in a format which allows the learners maximum opportunity to integrate the information. This presents a real challenge to any potential C^2 measurement method, as most existing manual (non-automated) methods are labor intensive and require the "clean-up" of raw data and the possible weighting or aggregation of several measures to develop composite performance scores.

Diagnostic

The requirement that C^2 measurement provide the information needed for diagnosis of individual/part-staff/whole-staff strengths and weaknesses is related to the requirement above regarding feedback. In order for the necessary corrective action to occur, the trainers not only must know whether performance is or is not deficient, they must also know by how much, and why it is deficient. Also, information concerning areas of strengths must be available in

order to provide reinforcing feedback. Global measures of team or individual effectiveness do not provide the necessary level of detail concerning performance to provide diagnostic feedback.

Discriminates

Discrimination in the context of measurement refers to the degree to which the measurement item or device identifies true differences among individuals/teams in regards to the behavior or attribute that the item is designed to measure.

The level of discrimination required for a given purpose is an issue which must be decided upon by the performance measurement developer. For instance, a measurement device which requires a dichotomous response (e.g., pass/fail or occurred/did not occur) can at most distinguish between two levels of performance. A seven-point rating scale can at most distinguish between seven levels of performance. In most measurement situations it is desirable to make the finest differentiations possible among levels of performance without sacrificing reliability.

Reliability

Reliability refers to the consistency or stability of measurement - for example, how consistent is the score of an individual from one time to another. Performance measurement indices provide a measure of performance at a particular time. Unless the measure can be shown to be reasonably consistent (that is, generalizable) over different occasions or over different samples of the same performance domain, little confidence can be placed in the results.

The concept of "reliability" requires that the purpose for which measurement is to be made be established in order to direct the focus of studies of the utility of the measurement. Measures of performance are neither reliable or unreliable in isolation. They are reliable (generalizable) over periods of time, over different samples of the behavior domain, over different raters, etc. It is possible for a measure to be more reliable in one of these respects than in another. The appropriate type of consistency in a particular case is dictated by the use to be made of the results.

A measure which provides inconsistent results cannot possibly provide valid information about the performance being measured. On the other hand, highly consistent results may be measuring the wrong thing. Thus, low reliability can be expected to restrict the degree of validity that a measure obtains. However, high reliability does not assure satisfactory validity. Reliability merely provides the consistency that makes validity possible.

Validity

Validity refers to the extent to which measurement results serve the particular uses for which they are intended. Basically, then, validity is always concerned with the specific use to be made of the results and with the soundness of the proposed interpretations of the results. Since validity is always specific to some particular use, it should not be considered a general quality. Measurement results are never just valid; they have a degree of validity for each particular interpretation to be made.

For C^2 diagnostic measurement, construct validity would be of key importance. Construct validity addresses the issues of whether an instrument measures what it was designed to measure, and how well it accomplishes that task. It provides the basis for interpreting a measure's results as a valid indicator of an individual's or unit's current status on the target task or construct. Empirical evidence of a measure's construct validity could be obtained by two principal methods in the context of C^2 training. One method would involve determining convergent and divergent correlational relationships of the measure with other measures. Evidence of convergent validity would be provided by relatively high correlations among those measures designed to assess a common, or related, construct, whereas evidence of divergent validity would require low correlations between the measure being assessed and measures designed to measure different, or unrelated constructs. For example, if an observational rating scale was developed to assess staff coordination during a C^2 training exercise, results of that measure should correlate higher with measures of tasks involving coordination, such as an information flow questionnaire (Kaplan, 1980) than with measures of tasks which do not require coordination, such as entry of incoming messages into the unit log, or other individual tasks. The second method of validation would involve examining the ways in which the measure behaves in regard to events occurring concurrently. It should show sensitivity to external variables which should impact upon the construct being measured. For example performance on most C^2 tasks could be expected to decline during periods of high stress. Therefore, the measure should show a decline of performance during the stress and recovery at termination of the stress.

Face validity should also be considered when designing a measure of C^2 training. Face validity is not validity in the technical sense, but pertains to whether the measure <u>appears</u> valid to the user. This is important because it may impact upon the acceptability of the measure to the user. If a measure is not well received by the users, the purpose of diagnostic performance measurement could be negated in that trainees may fail to internalize feedback derived from measures which appear to them to be irrelevant, inappropriate, or inapplicable. The developer of a measure can enhance face validity by being certain to formulate the measure in terms that appear relevant and plausible in the particular setting in which it will be used.

Ease of Administration

Ease of administration is concerned with the practical considerations involved in the implementation of performance measures. Measurement devices which have complicated directions, crucial timing aspects, requirements for complex apparatus, multiple observers, or extensive record keeping increase the possibilities of error when administered by individuals with little training or experience. These administration errors will, of course, have adverse effects on the validity and reliability of the measure. Furthermore, time available for performance measurement will always be at a premium in C^2 training situations, making it necessary to design measures which yield sound results in the shortest time possible.

Ease of Scoring

Ease of scoring is important for two reasons: results should be rapidly available for feedback, and less complicated scoring procedures generally produce more accurate results because of reduced opportunity for scoring error. Scoring can be achieved through direct or indirect means. Direct scoring requires little processing since the response is the result. For example, a checklist used to record the occurrence of an event would be scored directly. Indirect scoring requires that a response be transformed, aggregated with other responses, or analyzed in some way before scores are meaningful. Both types of scoring can yield useful data, but consideration should be given to simplicity of scoring.

Accurate

If a measure does not accurately measure the targeted performance, then interpretation and use of the results for feedback is worse than useless, and quite possibly harmful to the training efforts of the users. Unfortunately, the accuracy of measures for use in C^2 training is often very difficult to evaluate due to the lack of external criteria. This is also a problem for assessing the validity of instruments. The problem is exacerbated by the fact that high reliability can often be confused for accuracy. It is quite possible for a measure to yield reliable but inaccurate results, like a scale which always weighs a few pounds heavy, it is quite reliably inaccurate.

Objective

The objectivity of a measure refers to the degree to which equally competent observers, judges, test scorers, etc., obtain the same results through the use of that measure. That is, the results are not influenced by individual judgment or opinion. Objectivity is, of course, a matter of degree. A measure which is based upon judgment or opinion can be made more objective by providing a clearly specified criteria upon which to base judgments.

Automation Potential

With the increased use of computers in command group training it is desirable that performance measures have the realistic potential to cross over from manual to automated administration, data collection, and scoring. Automated measures would be more standardized and require fewer resources (time and personnel) to implement.

Sources and Dimensions of Performance Data

In C^2 training the principal sources of performance data are products, procedures, knowledge, decisions, and results. Table 1 shows the relationship of these sources to some of the dimensions of performance (timeliness, completeness, etc.), that they potentially produce. The list of performance dimensions is not intended to be an exhaustive inventory, but is merely offered as an example of the kinds of human performance dimensions one might examine in regards to the sources available. Furthermore, it is recognized that the sources of performance data (products, procedures, knowledge, decisions, and results) do not form clear, mutually exclusive categories, i.e., decision quality is affected by knowledge, products and procedures; results are related to decisions made and execution of procedures.

Table 1.

Performance Measurement Source/Dimension Matrix

	Sources of Performance Data				
	Products	Procedures	Knowledge	Decisions	Results
Dimensions of Performance					
Acceptable				X	X
Accurate	X	X	X		
Complete	X	X	X		
Consistent		X			
Efficient		X			X
Relevant	X				
Sufficient	X				
Timely	X	X		X	
Understandable	X		X		

Products

C^2 products are the formal outputs of the commander and/or staff tasks, such as estimates, plans, orders, reports, messages, logs, operations/situation maps, etc. In addition, products may exist in a concrete sense, such as a written order, or occur in verbal form only, such as a briefing or oral order. The dimensions of performance which could be derived from products would include timeliness, accuracy, completeness, understandability, relevancy, and sufficiency.

Procedures

C^2 procedures are established ways of executing tasks. Actions may become proceduralized through either tradition or Standard Operating Procedures, (SOP) and may involve either individual or team behaviors. Potential dimensions of performance derived from procedures are timeliness, accuracy, completeness, and consistency.

Knowledge

As a source of performance data, knowledge overlaps somewhat with products and procedures in that a body of relevant knowledge can be assumed to underlie product production and procedure execution. One must know what, when, and how to do before one can actually perform the required actions. However, it may be more efficient in some situations and for some types of knowledge to examine the knowledge base directly rather than the translation of that knowledge into behavior. Assessment of knowledge could also pertain to whether the C^2 trainees have an accurate picture of the ongoing battle. Examination of knowledge as a data source could derive the performance dimensions of accuracy, completeness and understanding.

Decisions

With the present state of the art in measurement, assessment of quality of decisions is still largely based on judgment, which requires speculation in regards to how different decision alternatives would have turned out, whether a decision differs from commonly accepted practice, whether a particular decision led to desirable battlefield results, and whether the decision was made with sufficient lead time to allow execution within the window of optimum opportunity. Therefore, the performance dimensions which could be derived from decisions are acceptability and timeliness.

Results

Results pertain to battle outcome or mission accomplishment. This source of performance data includes most of the standard operations research techniques for judging the outcomes of analytical wargames, such as loss-exchange ratio, (LERs), surviving maneuver force ratio differentials (SMFRDs), combat power ratios, etc., as well as indicators of the efficiency of an operation such as

consumption of supplies in various logistics categories (Solick and Lussier, 1986). The performance dimensions of efficiency and acceptability can be derived from results performance data.

Classes of Performance Measurement

For summarization purposes, existing performance measures can be categorized into three main classes: observation, testing, and statistical. Table 2 shows the relationship between these classes of performance measurement and the sources discussed in the previous section.

Table 2.

Performance Measurement Class/Source Matrix

	Classes of Performance Measurement		
	Observation	Testing	Statistical
Data Sources			
Products	X	X	
Procedures	X	X	
Knowledge	X	X	
Decisions	X		
Results	X		X

Observation generally relies upon some degree of human scrutiny and apperception, and may be accomplished by means of rating scales, checklists, note-taking, etc. The extent to which subjective judgment is a factor in this category of measurement varies according to the intent and design of the observations to be made, but some element of subjective impression is usually present. It is possible to capture performance data from any of the five primary sources through observation. The specific data captured depends upon the design of the rating scale or check-list. The three methods of capturing observation data which have been used in previous research are self-report, peer-report, and evaluation by outside sources. Three principal protocols currently exist for observation by outside evaluators: Army Training and Evaluation Program (ARTEP), Methodology for the Assessment of Planning Performance (MAPP) and the Headquarters Effectiveness Assessment Tool (HEAT).

Testing involves the direct assessment of performance by requiring the trainee(s) to engage in, or otherwise demonstrate, the behavior, knowledge, ability, etc., of interest, which is then scored or otherwise compared to the

expected or desired behavior. Testing approaches to measurement primarily capture performance data regarding products, procedures, or knowledge. The principal testing methods which have been used in previous research are information flow questionnaires (Kaplan, 1980); probes (Kaplan, 1979), and comparison of situation maps to ground truth.

Statistical measures yield performance data related to scenario outcomes and related measures. They are derived directly from the battle statistics (e.g., casualty rates or percentages, loss-exchange ratios (LERs) surviving maneuver force ratio differentials (SMFRDs), and combat power ratios). This category would also include statistical indices of the efficiency of an operation, such as consumption of supplies in various categories. These statistics are, of course, meaningless in themselves. They must be interpreted according to what would be expected of a unit with similar assets on a similar mission against a comparable opponent on similar terrain. The data which would make meaningful interpretation possible, however, do not currently exist (Solick & Lussier, 1986).

ASSESSMENT OF THE ADEQUACY OF EXISTING MEASUREMENT METHODS AS C^2 DIAGNOSTIC TOOLS

In this section, nine specific measurement techniques, which have been reported in the literature, will be analyzed using the measurement criteria presented in the previous section. These nine measurement techniques will be grouped according to measurement class (observation, testing, or statistical).

Observation

Self-assessment

Self-assessment is a subjective measurement method in which the individual is asked to evaluate himself. Self-assessment is believed by some to be a useful measurement tool since individuals have extensive data available about themselves and can provide insight that is not available from other sources. In addition, individuals generally attend to the situational factors which may impact upon their performance, whereas peers or outside observers may not be aware of, or take into account, such factors.

Available for Timely Feedback. Self-report evaluations usually employ simple rating forms or checklists which can usually be scored quickly and easily, or aggregated into a form usable for feedback to the training audience. Also, since the trainee is asked to rate himself in this measurement procedure, some feedback will occur intrinsically to the self-rating process.

Diagnostic. The utility of a self-report rating form or check list as a diagnostic tool would be dependent upon careful construction of the instrument. If scales are designed with tasks broken down into sub-tasks and behaviors so that it is possible to examine an audit trail to ascertain the origins and consequences of errors, then the scale will have diagnostic potential. No research report was available which specifically examined the diagnostic ability of self-assessment ratings.

Discriminates. Again, the ability of self-assessment to accurately discriminate differences in an individual's performance on different task dimensions is related to the construction of the instrument and the number of rating points provided.

Reliability. In regards to the consistency of self-assessment ratings over time, MacLane (1977, reported in Burnside, 1982) found that supervisors committed errors of inconsistency in 27 percent of their ratings, while the self-assessment inconsistency rate was only 9 percent. As noted earlier, individuals have extensive information about themselves which is not readily available to others. MacLane hypothesized that this enables individuals to support their judgment in regards to performance by examples of job related behavior, whereas

the supervisors in MacLane's study seemed to lack information about the individuals they rated and frequently could not support their appraisals with examples of behavior on the job.

Validity. Most research evidence concerning the construct validity of self-assessment ratings have compared self assessment to other subjective measures, with inconsistent results. Thornton (1980) reviewed studies which addressed this issue and found eleven studies which showed no relationship between self-appraisals and appraisals from supervisors or peers, while seven studies found at least a partial relationship between rating sources. No validity studies have been reported in which the validity of self-assessment has been examined against objective converging or diverging criteria, or in which self-assessment ratings were examined in regards to sensitivity to concurrent external variables.

Ease of Administration. Self-assessment ratings are relatively easy to administer since they do not require elaborate equipment or record keeping, however, they are resource intensive from the standpoint of the man hours required for administration. Since each individual completes his own assessment, the man hour requirements increase in proportion to the size of the training audience (i.e., a training audience of 10, each completing a self-assessment protocol requiring one hour, would result in 40 man hours of assessment time). In addition to the time resource requirement, there are potential difficulties in orchestrating the distribution and collection of assessment forms since members of the training audience are usually in several different geographical locations at the conclusion of an exercise.

Self-assessment could also be considered to be intrusive upon training, or at least upon training time, since self-assessment does not occur as part of the natural progression of the scenario and is not an integral part of the exercise.

Ease of Scoring. The ease with which self-assessment ratings can be scored is partially dependent upon whether direct or indirect scoring procedures are used. In either case, however, rating forms generally provide an instrument which is uncomplicated and simple to score.

Accuracy. As in studies regarding validity, studies which have investigated the accuracy of self-assessment have compared self-assessment with other subjective measures. Barber & Solick (1980) examined the ability of participants in a training exercise to rate their own performance. They found that, in general, participants tended to rate themselves higher than the ratings provided by external observers. However, since all ratings were subjective it is impossible to know which were more accurate. This finding is in agreement with other research concerning accuracy of self-assessment. Thornton (1980) reviewed the literature available which examined the accuracy of self-appraisal of job performance, with the conclusion that individuals rate themselves higher than they are rated by others. Self-ratings were shown to be higher than ratings by supervisors, peers, and assessment center raters. Meyer (1980) concluded, after examining years of related research, that most people have an

unrealistically positive perception of their job performance. He found that at least 40 percent of individuals typically rate themselves as being in the top ten percent in regards to performance, and that very few rate themselves below average. However, special measures can be taken to reduce the tendency of individuals to inflate their self-report ratings. For example, self-reports may be less lenient if the individual knows that self-assessments will be reviewed by a supervisor (Burnside, 1982).

Objective. Self-assessment depends almost totally upon subjective judgment. However, objectivity could be enhanced by providing clear criteria against which the individual can judge his performance.

Automation Potential. There appears little to be gained by automating self-assessment procedures in regards to improving efficiency or objectivity. However, electronic clipboards (Perceptronics, 1985) or some similar device could be used which would allow individuals to respond to self-assessment items which would then be automatically scored and/or aggregated into performance profiles. Normative data could be provided against which each individual could compare his scores or profiles.

Peer Assessment

Two methods of peer assessment are used most frequently: the rating procedure, where each member of a group rates every other member, and the nomination procedure, where each member of a group selects from the total group a given number of top and bottom individuals in terms of the attribute being evaluated.

In 1972, the office of the Deputy Chief of Staff for Personnel asked the U.S. Army Research Institute to investigate the value of peer evaluations in all officer schools, beginning with the Ranger Course. The data collected as a result of that effort provides the principal research evidence used in this assessment of the utility of peer assessments as a diagnostic evaluation tool.

Available for Timely Feedback. The research effort mentioned above did not directly address the question of the availability of peer reports for timely feedback. However, it was demonstrated that the scoring procedures could be adapted to a machine-processable optical scanning sheet, which should reduce the time required to process data for feedback (Downey, 1976).

Diagnostic. The utility of a peer rating format has diagnostic potential if the scale is constructed to provide such information. However, the peer nomination technique is not amenable to use as a diagnostic tool as this format yields one score which is interpretable as an individual's standing on the target global attribute.

Discriminates. This was not addressed directly by the research. The same principles of instrument construction discussed under "self-assessment" pertains to peer-assessment as well.

Reliability. Research conducted at the Ranger School using the nomination procedure resulted in high reliability coefficients across all phases of training (Downey, 1976). Split-half reliability coefficients ranged from .91 to .97, and test-retest reliability coefficients of .73 and .78 were reported. Peer assessment was also examined in regards to the selection for promotion of senior officers in 14 branches of the Army using the nomination procedure, which resulted in interrater reliability coefficients ranging from .63 to .94. The reliability of both peer ratings and peer nomination methods were examined with a sample of 125 Army officers attending Branch Basic Course. This study reported a split-half reliability coefficient of .90 for the rating method and .85 to .92 for the nomination method. Test-retest reliability for the rating method was .94, and .92 for the nomination method (Downey, 1974).

Validity. Using a sample of officers attending officer Branch Basic Course, Downey (1974) investigated the relationship of both peer ratings and nominations to scores on the Officer Evaluation Battery and grades in the Basic Course. This study indicated very small correlational relationships between peer ratings and nominations on the dimensions of the two external criteria. Validity coefficients for peer ratings ranged from .29 to -.36, and from .50 to -.32 for the peer nomination method.

In a study by Downey, Medland & Yates (1976) concerning a peer evaluation system for senior military officers, a point biserial correlational relationship of .39 was reported between peer nomination for promotion and actual promotion, with attendance at senior service college controlled.

In research on leadership effectiveness, Downey, Duffy, & Shiflett (1979) examined the convergent and divergent validity of peer assessment against a variety of measures from different sources. The peer evaluations showed little, or no, relationship to any of the converging variables and no evidence of appropriate divergence.

In work with peer assessment, Downey (1975) found acceptance by the military to be limited. No data was collected in an attempt to ascertain the specific nature of the acceptance problem, but one can assume that the peer assessment method suffered some degree of face invalidity for military users. In an attempt to enhance user acceptance, Downey (1975) studied the effectiveness of educating users about the utility of peer evaluations and stressing the importance of full participation. Results indicated that the educational and motivational treatment did improve user acceptance in that attitudes were moved from strongly negative to slightly positive.

Ease of Administration. Peer ratings and peer nomination methods are easy to administer since no elaborate equipment or record keeping is required. However, they have the same potential difficulties as self-assessment in regards to orchestration of administration to a geographically dispersed training audience. Peer assessment is also man-hour intensive and somewhat intrusive upon training.

Ease of Scoring. The ease with which peer ratings and peer nominations can be scored depends upon the method of scoring used. Peer ratings are usually scored directly, with the results pertaining to any one individual being the aggregation of his ratings across raters. Several different techniques exist for scoring peer nominations, some of which are rather complicated mathematically. In the study of peer assessment techniques at the Ranger School, Downey (1976) found that scoring could be facilitated with the use of a machine-processable optical scoring sheet.

Accuracy. In a summary of the research on accuracy of peer assessment, Burnside (1982) stated that peer assessments were more similar to supervisor appraisals than to self-assessments, but the relative accuracy of these approaches has not been adequately addressed. Kane & Lawler (1978) reviewed some of the related literature and concluded that no studies included an adequately objective measure of performance against which accuracy could be judged.

In the work by Downey (1976) investigating peer assessments for the Ranger School, the only criteria used for comparison with peer nominations which appear to be objective were Land Navigation total score, practical work exam, and patrol grades. It could not be definitely ascertained from the information provided in the report whether these criteria are objective performance scores or subjective evaluations. Nevertheless, peer assessments were found to correlate significantly in some instances with these criteria. However, the size of the relationship was moderate at best, since the highest coefficient obtained was .47.

Objective. Peer assessments are based upon subjective judgment only. The objectivity of peer rating formats could possibly be enhanced by providing clear criteria for rating. However, there is no obvious way to improve the objectivity of the peer nomination method.

Automation Potential. As with self-report, automating peer-report procedures could facilitate scoring and performance profile development.

External Sources

Three integrated protocols for observation by evaluators external to the training audience will be discussed. Army Training and Evaluation Program (ARTEP), Method of Assessing Planning Performance (MAPP) (Metlay, Liebling, Silverstein, Halatyn, Zimberg, & Richter, 1985), and Headquarters Effectiveness Assessment Tool (HEAT) (Defense Systems, Inc., 1984).

Army Training and Evaluation Program (ARTEP)

ARTEPs define the missions and tasks that are considered critical for a unit of a particular type and echelon level. The intention of ARTEPs is to describe the tasks to be completed, the combat condition under which the tasks must be performed, and the standard of performance which must be met. ARTEPs are in checklist format which observers score as "Go", "No Go", or "Not Observed".

Available for Timely Feedback. ARTEP guidelines call for each evaluator to orally critique the evaluated unit's strengths and weaknesses on the mission and/or tasks that he was assigned to evaluate. This is done soon after completion of evaluation. Each evaluator then provides a written explanation of weaknesses found. The senior evaluator consolidates feedback from all evaluators into a formal, written feedback package for the commander of the evaluated unit. This usually is not available for some time after completion of evaluation.

Olmstead, Baranick, & Elder (1978) developed a method using Brigade C^2 ARTEP tasks with a 7-point rating scale which resulted in a unit profile which graphically displayed the unit's relative strengths and weaknesses. The unit profile could be completed within approximately one hour by two people working together and could be used to provide feedback to the unit.

Diagnostic. The ARTEPs are only diagnostic in the sense that they provide information concerning whether or not evaluated tasks have been performed satisfactorily, and on the tasks which are not performed satisfactorily, they attempt to identify the section, leader, subunit, or other subgroup that failed to perform. Little is provided to explain why failure occurred or to what degree the performance was deficient.

In the method developed by Olmstead et al, the ARTEP ratings are compiled into a unit profile which permits comparison of scores on various tasks and identification of relative strengths and weaknesses among performance areas. However, no information is provided to explain why deficiencies occurred.

Discriminates. ARTEPs are checklists which evaluate observed performance dichotomously (Go, No Go), which permits discrimination at a very gross level.

The ARTEP method developed by Olmstead, et al, permits somewhat more discrimination since a 7-point scale is used.

Reliability. No reliability figures are available for true ARTEP performance data. However, several research studies have investigated the psychometric qualities of ARTEP tasks modified by the use of rating scales rather than dichotomous scoring. Kaplan & Barber (1979) investigated the C^2 ARTEP tasks evaluated on a 5-point scale to determine the desirability of this method of performance evaluation in C^2 training using battle simulations. The reliability estimates obtained in this study were quite low. Although no inter-rater reliability coefficient was reported, the report did state that "Individual raters differed in their judgment of subtask performance. The differences among ratings of the same command group by different observers were significant beyond the .001 level," (Kaplan & Barber, 1979, p. 45). With such a difference among raters, it can safely be assumed that interrater reliability was quite low. However, a later study by Thomas, Kaplan, & Barber (1984) which used the ARTEP tasks with a 9-point rating scale resulted in moderate levels of interrater agreement ($r = .63$). A third study (Thomas, Barber, & Kaplan, 1984) again investigated the use of ARTEP tasks, but employed a magnitude estimation scaling technique. (Magnitude estimation was used in an attempt to reduce

scale compression and ceiling effects, which are often found with rating scales that use a limited number of discrete categories, where raters tend to use only the upper part of the scale. In magnitude estimation, raters are asked to assess each subtask relative to a standard and assign a number to the subtask which reflects how many times greater or lesser it was than the standard.) This method of scaling ARTEP tasks also resulted in low inter-rater reliability ($r = .10$), with differences between raters being statistically significant beyond the .001 level.

Validity. The only validity figures available are from research investigating ARTEP tasks using rating scales. Given the low reliability figures reported, little confidence can be placed in the validity estimates obtained.

Thomas, Barber, & Kaplan (1984) examined the relationships of ARTEP ratings of performance to four simulation outcome measures, and reported negative correlation coefficients of low magnitude (-.04, -.24, -.06, and -.27). These correlation coefficients were not statistically significant.

Barber & Kaplan (1979) examined the relationship of several C^2 ARTEP tasks rated on a 3-point scale to other subjective ratings of performance effectiveness and mission accomplishment. Only two of the 33 correlations were significant at the .05 level or beyond. In addition, no consistency was found between raters or across time.

Ease of Administration. ARTEP evaluations do not require elaborate equipment. However, they are labor intensive and require substantial personnel resources to administer.

Ease of Scoring. Generally, ARTEPs are simple to score regardless of whether dichotomous or rating scale approaches are used.

Accuracy. The accuracy of subjective ARTEP evaluations has not been determined due to the lack of an independent objective criteria. However, the low reliability and validity estimates reported above would certainly limit the accuracy of ARTEP evaluations. Furthermore, research has indicated that ARTEP evaluators tend to use one general rating dimension (Medlin & Thompson, 1980) indicating an inability to differentiate among the dimensions of performance. A general impression of unit performance apparently is used to evaluate the unit, and more specific factors are used only if no strong overall impression is made. Appraisals of specific aspects of performance are unlikely to be accurate if based only upon general impressions.

Objective. ARTEP evaluations are based upon subjective judgment and are liable to the errors of perception which plague all subjective evaluations (for an overview of rater errors, see Garlinger, 1986). Although ARTEP purports to provide standards of performance to enhance objectivity of ratings, the standards for C^2 tasks are either non-existent or vague. However, those tasks which are evaluated dichotomously in regards to whether they occurred or did not occur are more objective than tasks calling for judgment regarding the degree of proficiency displayed.

Automation Potential. Performance evaluation of ARTEP tasks has the potential to be automated by presenting controllers or evaluators with the performance rating question(s) or checklist at the time the behavior occurs, or should occur. This would reduce the demands of memory in the rating task and eliminate the need for evaluators to take notes during the exercise. In the field environment the AkTEP evaluators could be provided with an electronic clipboard (hand-held computer device) on which the tasks to be evaluated and the performance criteria and rating scale appears on the screen. Date entry via touch-screen input would greatly simplify the paperwork aspect of such evaluations, (Perceptronics, 1985). Also, the Army Research Institute has developed a prototype system to computerize ARTEP production. The Computerized ARTEP Production System (CAPS) will support ARTEP authors by providing ARTEP data base storage, query and management as well as authoring and revision (Bloedorn, Crooks, Merrill, Saal, Meliza, and Kahn, 1985).

Method of Assessing Planning Performance (MAPP).

MAPP is a methodology developed at Hofstra University for the US Army Research Institute to evaluate performance of decision-making groups which employs both direct observation and the analysis of videotapes. The methodology was developed for use by the military in evaluating the decision-making process of command groups during the planning phase of training exercises. Seven phases of planning are defined and measured: information exchange, mission analysis, staff estimates, commander's estimate, preparation of plans, commander's approval, and operations order briefing. The methodology provides descriptions of these phases of the planning process and of activities which occur in each phase. Measurement scales were developed for each category of activity in each phase. The measurement scales are dichotomous (high/low or yes/no) on all items except two, which require time and frequency data.

Available for Timely Feedback. One of the criteria stated by the developers of the methodology was that it should be fast enough to generate data for feedback within 24 hours. However, no data have been provided as to the time required when the methodology was actually applied.

Diagnostic. MAPP is somewhat diagnostic in that it attempts to measure those behaviors which promote successful planning. However, the methodology is restricted to only one phase of training and does not provide information concerning why deficiencies occurred or to what degree deficiencies exist.

Discriminates. MAPP only discriminates dichotomously. Most items are scored either "high/low" or "yes/no".

Reliability. The only reliability estimates determined have been inter-rater reliability estimates for observations in the category pertaining to OPORD briefing. The inter-rater reliability for recording the duration of the individual briefings of the OPORD were .89 for one pair of observers and .94

for another pair. For observations concerning the presence or absence of target behaviors during the OPORD, the reliability coefficients were .99 and .90 for the two observation pairs.

Validity. No validity studies have been reported for this methodology.

Ease of Administration. The requirement to videotape the planning process for later analysis makes MAPP somewhat more cumbersome to administer than if the methodology relied solely upon direct observations.

Ease of Scoring. Items are scored dichotomously and require no further analyses or aggregation.

Accuracy. No information is available concerning the accuracy of MAPP.

Objectivity. The developers of MAPP state that it is an objective methodology. However, many of the items are scored "high/low" which requires subjective judgment by observers.

Automation Potential. Automation potential for MAPP would be similar to the potential for ARTEP evaluations discussed previously. Controllers or evaluators could be presented with the opportunity to rate performance at the time the behaviors should occur so as to reduce the demands on memory or note-taking, or videotaping.

Headquarters Effectiveness Assessment Tool (HEAT).

HEAT is an observational methodology for the assessment of headquarters performance and effectiveness developed by Defense Systems, Inc., (1984). HEAT can produce six overall measures and over a hundred other measures, which the HEAT developer refers to as diagnostic measures. The HEAT manual states, however, that a normal HEAT application would involve scoping the number of measures down to a reasonable and doable number. The methodology focuses on a HEAT model of the steps in the headquarters process: monitor, understand, consider alternative actions, plan, predict, decide, and direct. The six overall measures are derived from these six process steps. User involvement is required throughout the assessment process in determining what to measure, the standards against which the performance is compared, and in weighting the importance of individual tasks to the overall mission.

Available for Timely Feedback. Feedback of the results of a HEAT assessment are presented to the user in the form of a formal, written report, which takes several weeks to prepare.

Diagnostic. HEAT is diagnostic in that poor overall effectiveness scores can be linked with performance areas which contribute to the deficiency. Although the performance scores may provide information to pinpoint deficiencies, and comparison of performance scores to established standards will indicate the degree of deficiency, little information is provided concerning why deficiencies occurred. For example, one measure of performance is the percentage of

error in identifying strength of enemy units. A poor score on this item would certainly add insight to the reasons why a unit may have a poor overall effectiveness score, but would not contribute to understanding why the strength of enemy units was misjudged.

Discriminates. HEAT items are generally recorded as percentages (of units, time, etc), and so have greater discrimination power than if a few discrete categories were used for scoring.

Reliability. No reliability estimates of HEAT observation are currently available.

Validity. No validity estimates are currently available.

Ease of Administration. HEAT is an extremely complex methodology to administer. Five weeks of planning time are necessary to prepare for a HEAT assessment, observers are required and train-up time for observers takes several days. In addition, the actual implementation requires extensive record keeping and some data collection procedures have crucial timing requirements. Furthermore, there is anecdotal evidence that questions posed by data collectors to members of the training audience have been known to change the course of the exercise. This points to the need for an assessment of the training program provided for observers.

Ease of Scoring. Scoring of a HEAT application requires several weeks for specially trained analysts to complete. The data collected in one HEAT application were recorded on 12 different data sheets and calculated on 21 score sheets (DSI, 1984), providing multiple opportunities for scoring error.

Accuracy. No evidence is available concerning the accuracy of HEAT data.

Objective. Even though HEAT data are collected by observers, it is objective data (time, number, etc.) rather than data based upon judgment or opinion. However, due to the construction of data collection sheets, observers are frequently called upon to decide subjectively where to record an event and what events to record. Furthermore, the interpretation of the results is subjective in that there are no standardized performance criteria, each unit evaluated must establish its own criteria for success (Navy Personnel Research and Development Center, 1987).

Automation Potential. There is little obvious potential for automation of the HEAT methodology, with possible exception of using electronic clipboard data recording and automated aggregation and score development.

Probes

A probe is a method of controlling inputs, (information, events, requests, etc.), into a training exercise to elicit coordination, communication, and information processing behaviors within the command group. Probes have been used in manual training simulations for the purpose of exercising those staff areas, such as administrative and logistics functions, that were not well supported by the simulation. Probes can be combined with observation or other scoring plan, however, to provide a useful measurement technique (Solick & Lussier, 1986). The timing and content of probes is important as they should present situations and events which could be realistically expected to occur during an operation.

Probe objectives may include one or more of the following (Carter, Lockhart, & Patton, 1983):

- Analysis of the response behavior of a selected staff section in the performance of its functions.

- Analysis of the response behavior of the command group, including the commander and all staff sections in exercising command and control.

- Analysis of behavioral responses to a variety of stimuli.

- Analysis of different behavioral responses to a single stimulus.

Available for Timely Feedback. Results of probe measures can generally be promptly available for feedback since analysis requires only the comparison of the observed behavior to the expected behavior. Thomas, Kaplan, & Barber (1984) investigated the effect of feedback on probe performance and found that performance on probe measures significantly improved ($p < .05$) when probe pretest results were provided to the trainees as feedback.

Diagnostic. Probes designed to provide measurement of a specific objective or task can provide diagnostic information for that objective. Comparison of the obtained response to the appropriate or expected response could provide data concerning behaviors or actions which were omitted in the obtained response, or actions which occurred inappropriately in the obtained response.

Discriminates. Probe methodology has the potential to discriminate differences in level of performance on the target objective or task. The degree to which this potential is fulfilled is dependent upon the manner in which the obtained response is scored. However, method of scoring was not discussed in any of the published studies in which probes were used. If a dichtomous method of scoring was used (pass/fail or yes/no) then little discrimination would be possible. Fortunately, the probe technique appears to have the potential to be

scored in ways which enhance discrimination, such as a points system whereby points are awarded for each part of an expected response which occurs as part of the obtained response.

Reliability. No studies have been reported which investigated the reliability of probes. The reliability obtainable would be dependent upon the nature of the probe and the scoring method used.

Validity. No studies have been reported which investigated the validity of probes. However, a high degree of face validity can be assumed when probes are designed to elicit actions or behaviors which are recognized by the users as necessary and relevant.

Ease of Administration. Probes must be prepared and inserted into the exercise at the appropriate point. Preparation of probes requires careful planning so that the probe events appear realistic to the trainees. Furthermore, probe design must consider the scenario components (organization, mission, and environment) as a probe designed for one military operation may not be credible or useful for a different military operation. Therefore, the introduction of a probe into exercise play requires considerable knowledge by the controller responsible for insertion. In addition, once the probe is in play, controllers must be able to realistically respond to the ad hoc queries by the command group which the probe will generate, and recognize the events which should govern termination of probe play.

Ease of Scoring. No established method exists for scoring probes, but a number of possibilities are apparent. The expected response could be scored directly through observation in which a checklist of appropriate behaviors are assessed, or degree of performance proficiency could be assessed with a rating scale approach. Another possibility would be a point system in which points are awarded for the appropriate behaviors which are displayed. In any event, scoring of probes does not appear to present any significant problems which would make their use in C^2 diagnostic assessment infeasible.

Accuracy. Since the purpose of probes is to elicit specific behavior so that the behavior may be assessed, it may be more appropriate to evaluate probes in regard to their effectiveness in eliciting the target behavior rather than to evaluate accuracy. Thomas, Kaplan, & Barber, (1984) considered a probe to be effective if it was responded to on over 40% of the trials. A response rate of less than 40% indicated that the probe was too weak to generate reactions by trainees.

The method used to score the probes could be evaluated in regards to accuracy. However, no studies have been published in which probe generated performance scores were examined for accuracy.

Objective. Probe scoring methods may be either subjective or objective in approach. Performance indicators such as time to respond could easily be measured in an objective manner. Other indicators such as adequacy of response may be measured based upon subjective judgment of adequacy, or the response could

be compared to a criteria in which the percent of agreement is determined in order to achieve a more objective score. Many possibilities exist for achieving objective scoring of probe elicited behavior.

Automation Potential. Probe insertion into exercise play has potential for automation. However, automated measurement of the probe elicited response would be somewhat more difficult to achieve beyond a simple record keeping of whether or not anticipated behavior occurred, except for those probes which could be scored by matching response to the data contained in the tactical data system (TDS).

Information Flow Questionnaire.

This procedure for measuring information flow within the command group was reported by Kaplan (1980). The procedure requires that a multiple-choice questionnaire be administered to the members of the command group and company commanders at the completion of the planning phase of the exercise. The questionnaire measures recognition recall of specific items of information presented to individuals during the brigade briefing at the start of the planning phase. Taken as a whole, the questionnaire provides a measure of how well the people with information are disseminating that information to other individuals who are in need of it. Thus far, research with the information flow questionnaire has been limited to the planning and preparation phases of the battle since events are less predictable during the actual battle. However, it should be possible to develop information flow questionnaires to examine the flow of prewritten message information which is inserted into the exercise by controllers at the appropriate times.

Available for Timely Feedback. Thomas, Kaplan, & Barber (1984) found that when feedback concerning results of the information flow questionnaire pretest were provided to trainees, posttest results significantly improved. No mention was made in the published studies concerning the time required to score the questionnaire and aggregate data into a form usable and useful for feedback. However, it appears from the nature of the instrument that the information for feedback could be available in three to four hours.

Diagnostic. The information flow procedure provided data at three levels of the communication process: (1) communication from brigade to battalion, (2) communication within the battalion command group, and (3) communication from battalion to company (Thomas, Kaplan, & Barber). In addition, the performance of the command group can be further broken down to a communication matrix to examine intragroup communication channels, e.g., how much of the information required by the FSO from the S2 was actually received. These levels of analysis provide a desirable level of diagnostic potential in both pinpointing weak links in the communication process and in providing a communication audit trail. However, it should be noted that the methodology does not provide a means to distinguish whether an individual is not listening when information is presented, whether the information was not made available, or whether the individual does not recall the information correctly.

Discriminates. The information flow questionnaire employs an absolute scoring scale which provides a desirable level of discrimination when a sufficient number of items (> 5) are used.

Reliability. The only reliability coefficient reported for the information flow questionnaire was a split-half reliability of .82 reported by Thomas, Kaplan, & Barber (1984). The split-half coefficient is a measure of internal consistency (adequacy of item sampling), which provides no information concerning stability of the measure over time.

Validity. No studies concerning the validity of the information flow questionnaire have been published. However, validity would be suspect since variables other than those concerning the information flow process (memory recall, etc.), impact upon scores. Also, individuals may be able to recognize the correct answer on a recognition test but not be able to produce the correct answer through recall.

Ease of Administration. The information flow questionnaire is not difficult to administer and does not require excessive time for the trainees' responses, however, administration of the questionnaire is intrusive on training as it does not occur in the natural progression of the exercise and is not integral to the training exercise. Furthermore, a large amount of "front end" work is required as questionnaires must be tailor-made for each individual for each training scenario, as the information input pool cannot be standardized.

Ease of Scoring. Scoring of the questionnaire is simple and objective, each item is scored correct or incorrect based upon agreement with the appropriate response. Aggregation of scores to provide a diagnostic communication audit trail would be somewhat more complex, and would require analysis of items in individual questionnaires into reception and transmission scores for each individual with every other individual, resulting in a reception/transmission matrix.

Accuracy. No studies have been published which examined the accuracy of the information flow questionnaire. However, the same extraneous variables discussed concerning validity would impact upon accuracy as well.

Objective. The information flow questionnaire is a multiple-choice instrument with objective scoring procedures. Item responses are either correct or incorrect depending upon agreement with the keyed responses.

Automation Potential. Automated administration is the only obvious potential for applying automation to the information flow questionnaire.

Comparison of Staff Maps to "Ground Truth"

A comparison of the staff situation maps to the state of the real world battlefield is a measurement approach which assumes that the information posted on the various situation maps maintained by different staff sections is a reliable indicator of the state of knowledge held by the staff concerning the real world battlefield. (Solick & Lussier, 1986).

Available for Timely Feedback. When this approach was tried out in a student training exercise conducted at the Command and General Staff College (CGSC), one of the basic practical difficulties was obtaining the information posted on the maps and analyzing it in time to be useful for feedback (Solick & Lussier, 1986).

Diagnostic. Comparison of staff maps to ground truth can potentially provide diagnostic information concerning the currency of staff information. To be meaningful, however, performance standards must be established against which performance data can be compared (i.e., how much lag time must occur between actual events and reflection of that event on situation maps before it becomes a deficiency?) The degree to which this methodology can provide diagnostic information concerning the cause(s) of deficiencies in staff information would depend upon the data elements collected. For instance, if data elements were collected from both the situation maps and the staff logs, information would be available about whether deficiencies were observed because information was not received or because information was not posted to the map in a timely manner after being received.

Discriminates. This methodology has the potential to discriminate differences in levels of performance, depending upon development of a satisfactory method of scoring comparisons. Perhaps a point system could be developed in which points are scored for display of data elements within established tolerance levels of timeliness.

Reliability. Due to the practical difficulties encountered in collecting and scoring performance data using this technique, no studies have been conducted to examine reliability. Reliability would ultimately depend on the nature of scoring methods developed.

Validity. The validity of performance measures derived from comparing situation maps to "ground truth" has not been examined. However, face validity can be assumed, as the maintenance of staff maps is generally recognized by the military community as being necessary and relevant.

Ease of Administration. Previously attempted manual methods of collecting the information reflected on the situation maps is cumbersome and impractical. When this method was implemented in the CGSC training, the attempt was made to capture the information contained on the maps by photographing these maps periodically. In computer driven exercises, obtaining the "ground truth" information for comparison is somewhat easier as this data can be obtained from computer printout.

Ease of Scoring. Two sets of data are required in order to derive performance measures from this methodology. One set consists of computer printouts representing the actual state of the battlefield. The other set represents the state of the battlefield as represented by the situation maps. To derive performance measures from these two data sets, they must be be translated into a common format, and the analysis must focus on a limited subset of the information available in the two data sets so that the battlefield time represented by the photographs or other representation of situation maps can be matched with the appropriate section of computer printout of ground truth. No simple or easy method currently exists for accomplishing these tasks.

Accuracy. Accuracy of this methodology has not been examined. However, accuracy would ultimately depend upon method of collecting and scoring data.

Objective. The objectivity of this method would ultimately depend upon the manner in which comparisons are made between the situation map data set and the ground truth data set.

Automation Potential. Automation appears to be the best hope for the development of this technique into a practical tool for measuring performance. This would involve a computer comparison of the "ground truth" in the model with the information "posted" by the staff to their tactical data system (TDS) (Solick & Lussier, 1986).

Statistical

Results Data

Results data examined in the past have consisted primarily of various methods of computing battle outcome ratios and measures of effectiveness such as fuel or other resource consumption.

Thomas & Cocklin (1983) and Thomas, Barber, & Kaplan, (1984) examined various ways of combining friendly and OPFOR strength and losses to portray the degree of success of friendly forces in simulated combat. These measures are presented in Table 3. Relative Exchange Ratio (RER) and Surviving Maneuver Force Ratio Differential (SMFRD) are modified versions of indices obtained from combat development studies (USACDC, 1973).

RER is simply the ratio of the proportion of OPFOR losses to the proportion of friendly losses. SMFRD is calculated by subtracting the proportion of OPFOR surviving battle from the proportion of friendly forces surviving. Both measures are, therefore, rather straight-forward comparisons of losses or surviving strengths of opposing forces, (Kaplan, 1985).

Table 3

Definitions of Selected Simulation Outcome Measures

Measure	Definition
RER =	$\frac{\text{OPFOR Losses/OPFOR Initial Strengths}}{\text{Friendly Losses/Friendly Initial Strengths}}$
SMFRD =	$\frac{\text{Friendly Remaining Strength}}{\text{Friendly Initial Strength}} - \frac{\text{OPFOR Remaining Strength}}{\text{OPFOR Initial Strength}}$
C^2ILL =	$\frac{1}{2}\left\langle\frac{\text{Friendly Remaining Strength}}{\text{Friendly Initial Strength}}\right\rangle + \frac{\text{OPFOR Losses}}{\text{OPFOR Initial Strength}}$
ΔCR =	$\frac{\frac{\text{OPFOR Initial Strength}}{\text{Friendly Initial Strength}} - \frac{\text{OPFOR Remaining Strength}}{\text{Friendly Remaining Strength}}}{\frac{\text{OPFOR Initial Strength}}{\text{Friendly Initial Strength}}}$

* Taken from Thomas, Barber, & Kaplan (1984).

The Command and Control Index of Lethality Levels (C^2ILL) is based on the assumption that it is preferable to have a high percentage of forces surviving, while attriting a relatively high proportion of enemy forces. Hence, C^2ILL is computed by adding the two components together. The proportion of friendly forces surviving is divided in half because it was observed that in covering force missions, controller ratings of performance were more responsive to the amount of enemy forces attrited than to the amount of friendly forces surviving. The weighting factor places a higher emphasis on OPFOR losses (Kaplan, 1985).

The change in Combat Ratio (Δ CR) is based on the assumption that it is preferable to end a battle with a higher combat ratio than existed prior to battle. The measure, therefore, computes the change in combat ratio relative to initial combat ratio, (Kaplan, 1985).

In all the above measures, higher values indicate greater success for the friendly forces. All measures of initial strength and losses were based on equipment and not personnel. All types of combat equipment were considered in the calculations, where tanks, APC's, TOWs, etc., were combined. Combination

was achieved by adding the products of each equipment type and its corresponding combat effectiveness weight. The CATTS battle calculus included combat effectiveness weights for each piece of equipment based on its ability to destroy other types of equipment, e.g., an M60A1 tank had a weight of 73, an M113 APC a weight of 19, and a T62 a weight of 80, (Kaplan, 1985).

Available for Timely Feedback. Measures of outcome and mission accomplishment can probably be generated quickly in computer driven training exercises.

Diagnostic. Statistical outcome measures offer little of value in diagnosing performance strengths and weaknesses as there is no measurement of the tasks performed by the command group, how well these tasks are performed, or the way in which they are performed. While it may be reasonable to assume that the command group is a necessary element for desirable battlefield outcomes, it also is evident that the performance of the command group is not sufficient to insure success. Battle scenario characteristics, missions, performance of friendly and OPFOR controllers, and data entry personnel are but a few of the potentially significant uncontrolled variables impacting upon outcome statistics.

Discriminates. Differences in levels of performance of individuals within the command group cannot be distinguished by statistical outcome measures in as much as they are global indices of whole group performance. Furthermore, given the abundance of uncontrolled variables which impact upon these measures, attempts to use them to discriminate levels of group performance is hazardous.

Reliability. The raw numbers which provide the data for computation of battle outcome measures are probably quite reliable, especially in computer driven training exercises in which the generation of the data can be accomplished through automation. However, no studies have been reported in which the reliability of battle outcome measures has been examined. Perhaps this is because such a task would require that the uncontrolled variables impacting on these measures be controlled or held constant across trials.

Validity. Studies which have examined the validity of battle outcome measures have provided conflicting results. Thomas (1983) investigated the validity of the battle outcome measures listed in Table 3 by correlating these measures with controller ratings of performance for three types of units (Mech, Infantry, Cav). Each outcome measure correlated significantly with performance ratings for one or more of the unit types, but not for all three. These findings appear to suggest that the validity of a particular outcome measure may be unit dependent. Thomas further examined the relationship of battle outcome measures to performance ratings with mission held constant. This analysis resulted in significant correlations for RER, SMFRD, and C^2ILL across all mission types and units. The highest correlations obtained were for C^2ILL and SMFRD, although no outcome measure appeared to be the "best" predictor of controller ratings in all situations; however, Thomas concluded that C^2ILL appeared to be the most consistent.

Further research by Thomas & Cocklin (1983) indicated that SMFRDs were the only battle outcome measures which accounted for a significant portion of the variance in ratings of mission accomplishment provided by a panel of military experts. Regression modeling was used to derive an optimal weighted linear combination of SMFRDs, measures of territory lost, time the enemy was delayed and the accuracy of intelligence estimates which accounted for an average of 98% of variance in the mission accomplishment judgments of the panel of military experts.

A study by Thomas, Barber, & Kaplan (1984) found that none of the battle outcome measures correlated significantly with controller performance ratings.

Ease of Administration. When battle outcome measures are derived from computer driven training exercises, no administration problem exist as the data can be obtained from computer records. However, collection of data to calculate battle outcome scores when the training is conducted with manual simulation in the field environment could be quite cumbersome, requiring large personnel resources and record keeping activities.

Ease of Scoring. Battle outcome measures are numerical ratios which require no excessive computational abilities or resources given that a computer model provides the necessary data.

Accuracy. The accuracy of battle outcome measures depends upon the accuracy of the underlying models of combat which provide the attrition results. Considerable doubt exists as to the accuracy of current models (Solick & Lussier, 1986). This is particularly true when firepower score methods are used to estimate the relative ability of a unit to inflict OPFOR casualties. This distrust results from the failure of firepower scores to consider the differential effectiveness of various weapon systems against targets of varying "hardness." Better attrition methodologies are still suspect when used in conjunction with unclassified weapons effects data (Solick & Lussier, 1986).

Objective. Battle outcome scores derived from computer driven exercises in which the raw data collection is automated will be somewhat more objective than scores derived in a field training environment in which data collection is accomplished manually and relies on human observation and judgments concerning attritions.

Automation Potential. Outcome/mission accomplishment measures are currently automated in simulation systems. The automation challenge for such measures lies with the development of automated analytical models to make it possible to standardize outcome measures to a sufficient degree to permit the collection of a normative data base which can make interpretation of outcome measures possible.

POTENTIAL APPLICATION OF OTHER MEASUREMENT METHODS AS C^2 DIAGNOSTIC TOOLS

There are other behavioral measurement techniques worth considering as options for diagnostic assessment. Some are predecessors or variations of those reviewed to this point in this paper. Others have not been applied in CGT or C^2 exercises. A brief description of some of these techniques follows, including a discussion of their strengths and weaknesses. A systematic review of each technique against the ten criteria is not appropriate since little or no data have been collected in the CGT environments and specific details of applications to those environments have not been worked out. For a more comprehensive review of many of these techniques the reader is referred to Meister (1985).

Under observational techniques, one which is similar to self-assessment is a technique founded in the cognitive sciences. Introspection or the think aloud protocol is an individual observation technique used to collect information on internal thought processes of an individual. The purpose of the technique is typically descriptive in nature rather than evaluative. It could be used to address procedures, knowledge, and possibly decisions. As the technique requires the individual to provide a commentary on some manner of his behavior, it can interfere with performance and may alter the behavior as the individual is required to consciously think about and express what he is doing. Verbal protocols would be disruptive to other members of the staff.

The critical incident technique is another observational technique, also descriptive in nature, which primarily focuses on procedures. It attempts to infer those behaviors which relate to successful performance. Relying on observation by a dedicated observer, key incidents are noted which appear to have a critical impact on system performance. It is most suitable to situations where results are clearly observable or defined and where there are direct relationships with behaviors. Usually, however, a clear link between procedures and battlefield results is not the case in staff performance.

Time and motion analysis is yet another observational method for examining procedures. A number of variations have been used, but the common theme among them is emphasis on psycho-motor processes which is beneficial for analyzing manual assembly work. For CGT it is not a useful or practical technique to examine processes at such a micro-level, but it is worth mentioning that several different sampling variations have been used in time and motion analysis. One sampling approach is to use observation at fixed intervals of fixed duration. Another is to use various intervals. In yet another sampling approach, observation is triggered by the occurrence of a given event. These variations in sampling also have been applied to other measurement schemes.

From the equipment/system design discipline, a number of analytical techniques have been developed for describing and evaluating procedures, equipment layout, personnel task allocation, and human reliability diagnosis. Among these techniques are link analysis, operational sequence diagrams, and decision/action/information diagrams. Although these analytical techniques are not

measurement techniques in the same sense as the others which have been discussed, they can be used in CGT to analyze and structure task processes, to set normative standards for performance, and to organize data collection.

A measurement area which is so general that there is no special name associated with it, is the collection of time, accuracy or frequency data on task performance. These data may be gathered in many ways, ranging from external observers recording data onto a clipboard to automatic data recording when the staff member is using an automated tactical data system. Task data may be of as many types as there are tasks and criteria dimensions. Any objective measurement technique will relate undoubtedly in some manner to time, accuracy, or frequency.

In addition to the task process, the tangible products from those tasks can be assessed. The evaluation of staff reports and orders can be done to determine their timeliness, completeness, and quality. The products can be rated against some established standards, such as from doctrine or standard operating procedure (SOP), and/or be judged by subject matter experts. One difficulty is that a poor or good rating of a product does not correspond necessarily to a poor or good rating of staff procedures, knowledge, decision quality, or battlefield results.

Another area, which is not a measurement technique per se, but provides an organized body of literature in which measurement is a key issue is workload analysis. Though physical workload limits are a consideration, the central concern is mental workload, i.e., can an operator manage an acceptable level of performance output under periods of high workload. Measures to address this and corollary sets of issues come from four areas: physiological measures, primary task performance, secondary task performance, and subjective assessment. The second and fourth areas are not different from those techniques already tried in CGT. Physiological measures are probably too intrusive and at too basic a level of examination to provide any near-term measurement potential for CGT. Secondary task performance may have some usefulness as a CGT technique, especially as a specific training objective approximates the workload issue, (e.g., to perform some task at some minimal level of performance under some level of workload, perhaps defined as the number of incoming messages or reports per time interval). Secondary "tasks are similar to the concept of probes. Unlike the typical use of probes, secondary tasks are in a unrelated task domain, (e.g., Sternberg, 1969). The technique is intrusive, is at an individual performance level, and requires advanced preparation and training. However, it also offers standard task stimulis," and it is selected or developed based on the ease of information and quantification.

Under the testing class of measurement, a technique similar to the information flow test is a written proficiency examination. Whereas the information flow addresses the knowledge of the specific tactics and operations occurring in an exercise, a general knowledge test would assess knowledge of staff procedures, responsibilities, and doctrine. Although this is a basic educational approach to measurement of knowledge it has not been attempted in CGT until

recently, when pre and post training versions of a proficiency test were developed to evaluate Army Training Battle Simulation System (ARTBASS) training (Garlinger, Fallesen, Solick, & Lussier, 1986). While it may be an appropriate evaluation technique, unless the test is tailored to specific exercises it will not provide much specific material for diagnostic feedback.

A technique which may provide insight into performance diagnosis, but which has not been tried, would be the collection of staff perceptions. This technique would primarily address knowledge and its application. Using a testing or observation approach, beliefs about enemy intentions, friendly status, resources available, etc., could be collected to compare with actual status. Instead of using ground truth as the comparator, the actual status limited to only that information which was available to the staff would also be appropriate.

An area originating in communications work is an information theoretic approach to performance assessment. The most commonly applied level of the theory involves the measurement of the amount of information transmitted or acquired. The quantification is based on the probability of receipt of a particular message. The usual application in psychology has been the study of the effects of intervening variables on the perception of information. Of greater significance are the two higher levels of Shannon and Weaver's (1949) theory: (a) the meaning of the transmitted information, and (b) how the information is used, once it has been received and understood. Only a small amount of basic measurement development has occurred in this area and it is not of sufficient maturity to apply to CGT diagnosis and feedback.

Measuring how effectively information is used based on the course of action selected for a decision problem addresses decision quality. No techniques for assessing decision quality have been found acceptable for CGT because of the difficulty in ascertaining the results of a command decision in terms of battlefield effectiveness. To do this requires some way of knowing (or predicting) what cause-effect relationship is in effect between a decision and the resulting outcome for a given situation and knowing the effects of alternate decision options which were not selected nor executed.

One final note is that no one combination of class of measurement with data source will provide complete measurement for all CGT objectives on which feedback is desired. Individual measurement techniques need to be matched to specific task training objectives. Even then, to address a task, several measures and possibly several techniques will be needed to obtain the requisite data. A figure of merit can be used to put the diagnostic results into a summary form, combining results of measures to give an overall score of "goodness". A figure of merit is subject to extensive development and validation work involving what measures to include, what weighting scheme to use among component measures, and making it robust over the range of conditions employed in command group training.

SUMMARY AND CONCLUSIONS

Of those measurement techniques which have been investigated and reported in command group training applications, there has been no overwhelming "success". None of the nine techniques favorably met all ten measurement criteria specified for this review. The difficulty in selecting or developing a staff measurement technique comes from the complexity and diffusion of C^2 tasks. The questions of what should be trained (training objectives), how good does performance need to be (performance standards), what to measure (task and behavioral variables), and how to measure (measurement technique) are some of the issues which must be resolved before measurement tailoring guidelines are developed.

Techniques in the observation, testing, and statistical classes were found to have a variety of advantages and disadvantages. Self and peer assessment are favorable in terms of availability for timely feedback, reliability and ease of scoring, but fail to be acceptable in terms of objectivity and accuracy. ARTEPs are an in-place technique, but have questionable reliability, validity, and accuracy as they exist in present form. The MAPP is an immature technique which does not fare well in most of the criterion categories. The HEAT technique is a cumbersome method to administer and score and has unknown reliability, validity, and accuracy. The probe technique appears to have potential to meet the criterion categories for which currently there is no information. However, probes are a prompting or sampling (on-occurrence) technique with no inherent method for data collection. The information flow test fares well in most categories, however it is limited primarily to communication tasks (information acquisition and dissemination), and the resulting scores are dependent on memory recall. The staff map assessment has failed to work well in application, but has good potential to be automated. Summary techniques such as results data are not diagnostic and have questionable accuracy when it comes to assessing staff procedures. The other potential techniques undoubtedly would have the same types of disadvantages as those techniques which have been investigated. From this review of C^2 staff measurement techniques, it can be concluded that no one technique is acceptable in its present form for training diagnosis and feedback, and that some combination of techniques with refinements will be needed.

The purpose of this measurement review has not been to document the development or selection of a measurement technique, rather it has been a study of the strengths and weaknesses of existing and potential techniques. However several conclusions based on the review are appropriate:

a. External observers are to be preferred over peer or self assessment.

b. Probes can enhance a training exercise as well as present situations for measurement of subsequent performance.

c. Information flow and other testing techniques rate better than observation or summarization techniques in terms of objectivity, accuracy, validity, and reliability.

d. A diagnostic assessment and feedback system should rely on multiple measurement techniques in order to be able to "diagnose" cause and effect relationships and to address potential training objectives.

e. Training objectives concerning practice on tasks or acquisition of skills, if based on a structured hierarchial format, will enhance the matching of measurement techniques to tasks or skills.

Future Research Directions

None of the reviewed techniques which have been applied to CGT had documented information on all ten of the identified criteria, so determining their suitability is difficult. Developers and practitioners involved with CGT measurement need to attend to these or selected criteria and attempt to verify the adequacy of the techniques as opportunities arise. Better information on these criteria needs to be obtained for the most promising of techniques. Verification designs which occur naturally in training and the process of data collection need to be recognized and learned from. Where possible, specific designs and supporting data need to be used to obtain vital information on accuracy, validity, and reliability. Multiple measurement approaches need to be tried simultaneously. Since specifying a criterion measure for command group performance has been unsuccessful, the determination of candidate measures' validity and accuracy has been limited. Verification approaches, to include convergent validity and sensitivity analyses of the multiple measures, need to be used more often to enable the assessment of the techniques' suitability.

New measurement approaches must be considered and evaluated to determine if other techniques are more appropriate than existing ones for providing data on training diagnostics and feedback. Increased effort is going into introspective techniques because of the crucial role they play in knowledge engineering. If introspection becomes more formalized, it may be used in command group training for the purpose of obtaining information on perceptions of staff members and the intermediate steps in their decision making processes. It is important that verifying techniques be tried to confirm that introspection provides accurate information.

Improved techniques are needed to measure how information is used. Research on information usage should pursue quantifying the selection, interpretation and use of information as it occurs to affect an intended result. Existing techniques which do this are limited to simpler, more constrained situations than those which are faced by the staff. Research and development is required to extend the measures to more real-world settings and problems.

A third line of research should examine the suitability of a secondary task approach for command group training applications. A task or set of tasks would need to be selected or developed to stimulate staff activity, much as individual probes do. By having a secondary task that is more readily observed and measured than many of the tasks required of the staff, primary task performance can be scored in terms of efficiency and spare capacity for the secondary task.

Solick and Lussier (1986) have made recommendations for development and research on seven measurement techniques which can be implemented in CGT simulations:

- Develop templates for instituting the information flow methodology, to be filled in with specific items of information from the scenario. Document the methodology and the sources of information in the data base that are to be used.

- Develop a list of probes, along with a means for automatically notifying appropriate controllers to insert them, either at pre-set times or in response to simulation events.

- Provide automatic detection of events based on common errors that indicate failures in staff planning or coordination.

- Develop normative data from model runs for interpretation of mission accomplishment data.

- Develop a watchdog program for the staff's tactical data system to track preparation and delivery of reports.

- Develop procedures to compare the ground truth data in the training system data base with the staff's picture of the battle as reflected in the tactical data system.

- Develop analytical wargaming procedures to evaluate alternative decisions.

These recommendations address both the general problem of lack of objective measurement being used in CGT and the specific weakness of various techniques which have been identified in this review. These are all techniques which are rich in issues for research and development.

Most importantly, better modeling and analysis are needed of the behavioral aspects of the command and control environment and behaviors that take place in it. Of course if there were better measurement techniques then there would be a better understanding of the behaviors. Future research and analysis should emphasize the increase in differentiation on each different dimension of performance and among the different dimensions to assure that the correct dimensions of behavior are being diagnosed for training feedback. A key to better performance measurement is the exploration of the relationships among the sources of performance data. By collecting different sources of performance

data, critical questions can be addressed which could help determine what sources of performance data should be of interest for which training applications. Example source-related questions include how does good "knowledge" affect the quality of "products," to what extent do variants in "procedures" affect "decisions," how do metrics of "decision" quality relate to battlefield "result" scores, what are the key situational variables? Research conducted to collect measures from the different sources of performance will provide a better basis of understanding of the command and control process and a better repertoire from which measures can be selected.

References

Anastasi, A. (1982). Psychological testing. New York: MacMillian Publishing Co. Inc.

Army Science Board (1985). Final Report of the 1985 Summer Study on Training and Training Technology - Applications for AirLand Battle and Future Concepts. Washington, D.C.; Office of the Assistant Secretary of the Army (Research, Development, and Acquisition). DTIC AD B101 040 L-2.

Barber, H. F., & Kaplan, I. T. (1979). Battalion command group performance in simulated combat Technical Report 353. Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

Barber, H. F., & Solick, R. E. (1980). MILES training and evaluation test USAREUR: battalion command group training Research Report 1290. Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

Bloedorn, G. W., Crooks, W. H., Merrill, M. D., Saal, H. J., Meliza, L. L., and Kahn, O. I. (1985). Concept study of the computer-aided ARTEP production system (CAPS). Research Report 1403. Alexandria, VA: U.S. Army Research Institute for the Behaviorl and Social Sciences.

Burnside, B. L. (1982). Subjective appraisal as a feedback tool Technical Report 604. Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

Carter, C. F., Jr., Lockhart, D. C., & Patton, M. S. (1983). Command group behaviors: their identification, quantification, and impact on collective output in automated and non-automated environments Annual Technical Report. Science Applications, Inc., 424 Delaware, Leavenworth, Kansas

Defense Systems, Inc. (1983). Theater headquarters effectiveness: its measurement and relationship to size, structure, functions, and linkages. Report prepared for C^3 Architecture and Mission Analysis, Planning and Systems Integration Directorate, Defense Communications Agency.

Defense Systems, Inc. (1984). HEAT's user's manual. Prepared for Defense Communications Agency under Contract No. DCA-100-84-C-0047.

Downey, R. G. (1974). Associate evaluations: nominations vs. ratings Technical Paper 253. Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

Downey, R. G. (1975). Associate evaluations: improving field acceptance Research Memorandum 75-5. Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

Downey, R. G. (1976). Associate nominations in the U.S. Army officer training environment: The ranger course. Research Problem Review 76-8, Alexandria, VA: US Army Research Institute for the Behvioral and Social Sciences.

Downey, R. G., Duffy, P. J., & Shiflett, S. (1979). Construct validity of leader effectiveness criteria Technical Paper 368. Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

Downey, R. G., Medland, F. F., & Yates, L. G. (1976). Evaluation of a peer rating system for predicting subsequent promotion of senior military officers, Research Memorandum 76-7. Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

Downs, C. W., Johnson, K., & Barge, J. K. (1984). Communication feedback and task performance in organizations: A reveiw of the literature. Organizational Communication, 9, 13-47.

Dyer, R., Matthews, J. J., Stulac, J. F., Wright, C. E., & Yudowitch, K. (1976). Questionnaire construction manual, annex, literature survey and bibliography. Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences (ARI BSS p-77-2).

Gade, P. A., Fields, A. F., & Alderman, I. N. (1978). Selective feedback as a training aid to on-line tactical data inputting. (Technical Paper 349) Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

Garlinger, D. K. (1986). The effectiveness of a rater training booklet in increasing accuracy of performance ratings. Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences, Research Note in press.

Garlinger, D. K., Fallesen, J. J., Solick, R. E., and Lussier, J. (1987). Appraisal of Army training battle simulation system (ARTBASS) training: Test validation phase. Technical Report in press. Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

Gilbert, A. C. F., & Downey, R. G. (1978). Validity of peer ratings obtained during ranger training Technical Paper 344. Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

Ilgen, D. R., Fisher, C. D., & Taylor, M. S. (1984). Performing feedback: A review of its psychological and behavioral effects. (Research Note 84-47) Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

Kane, J. S., & Lawler, E. E. (1987). Methods of peer assessment. Psychological Bulletin, 85, 555-586.

Kaplan, I. T. (1980). Information flow in battalion command groups. Technical Report 499. Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

Kaplan, I. T. (1985). Lessons learned in research on command group training Technical Report in press. Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

Kaplan, I. T., & Barber, H. F. (1979). Training battalion command groups in simulated combat: identification and measurement of critical performance Technical Paper 376. Alexandria, VA: US Army Research Institute for the Behavioral Sciences.

MacLane, C. N. (1977). Promotion evaluation for inter-organizational referral: A behavioral expectation approach. Paper presented at the Military Testing Association Conference, San Antonio, October.

McIntrye, R. M., Smith, D. E., & Hassett, C. E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. Journal of Applied Psychology, 69, 147-156.

Medlin, S. M., & Thompson, P. (1980). Evaluator rating of unit performance in field exercises: A multidimensional scaling analysis. Technical Report 438. Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

Meister, D. (1985). Behavioral analysis and measurement methods. New York: John Wiley & Sons.

Meyer, H. H. (1980). Self-appraisal of job performance. Personnel Psychology, 33, 291-296.

Metlay, W., Liebling, D., Silverstein, N., Halatyn, A., Zimberg, A., & Richter, E. (1985). Methodology for the assessment of the command group planning process. Unpublished research report. Applied Research and Evaluation Program, Hofstra University.

Navy Personnel Research and Development Center (1987). Evaluation of the headquarters effectiveness tool: Defense Systems, Inc., battle force impact training command and control evaluation guide. San Diego, CA.

Nunnally, J. C., & Wilson, W. H. (1975). Methods and theory for developing measures in evaluation research. In Struening, E.L., & Guttentag, M. (Eds). Handbook of Evaluation Research, Vol I. Beverly Hills, CA: Sage Publications, Inc.

Olmstead, J. A., Baranick, M. J., & Elder, B. L. (1978). A training feedback system for brigade command groups Technical Report 78-A19. Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

Olmstead, J. A., Elder, B. L., Forsyth, J.M. (1978). Organizational process and combat readiness: feasibility of training organizational staff officers to assess command group performance Technical Report 468. Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

Perceptronics (1985). The electronic clipboard system. (Videotape) North Hollywood, CA: The Video Tape Company.

Shannon, C. E., & Weaver, W. (1949). The mathematical theory of communication. Urbana, Illinois: Universtiy of Illinois Press.

Solick, R. E., & Lussier, J. W. (1986). Design of battle simulations for command and staff training Technical Report in press. Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

Sternberg, S. (1969). The discovery of processing stages: Extension of Donders' method. Acta Psychologica, 30, 276-315.

Thomas, G. S. (1983). Battle simulation outcomes as potential measures of BCG performance in CATTS exercises Working Paper FLvFU 83-1. Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

Thomas, G. S., Barber, H. F., & Kaplan, I. T. (1984). The impact of CATTS system characteristics on selected measures of battalion command group performance, Technical Report 609. Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

Thomas, G. S., & Cocklin, T. G. (1983). A model of mission accomplishment in simulated battle Technical Report 599. Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

Thomas, G. S., Kaplan, I. T., & Barber, H. F. (1984). Command and control training in the combined arms tactical training simulator Technical Report 615. Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

Thorndike, R. L. (1982). Applied psychometrics. Boston: Houghton Mifflin Company.

Thornton, G. C. (1980). Psychometric properties of self-appraisals of job performance. Personnel Psychology, 33, 263-272.

USACDC (1973). Force development: the measurement of effectiveness. USACDC Pamphlet 71-1.

Zedeck, S., & Cascio, W. F. (1982). Performance appraisal decisions as a function of rater training and purpose of appraisal. Journal of Applied Psychology, 67, 752-758.